HEAD-TRACKING AND SUBJECT POSITIONING USING BINAURAL HEADSET MICROPHONES AND COMMON MODULATION ANCHOR SOURCES

Matti Karjalainen, Miikka Tikander, and Aki Härmä

Helsinki University of Technology
Laboratory of Acoustics and Audio Signal Processing
FIN-02015 HUT, Finland
{firstname.lastname}@hut.fi

ABSTRACT

A prerequisite in many systems for virtual and augmented reality audio is the tracking of a subject's head position and direction. When the subject is wearing binaural headset microphones, the signals from them can be cross-correlated with known sound sources, called here anchor sources, to obtain an accurate estimate of the subject's position and orientation. In this paper we propose a method where the anchor sources radiate in separate frequency bands but with common modulation signal. After demodulation, distance estimates between anchor sources and binaural microphones are obtained from positions of cross-correlation maxima, which further yield the desired head coordinates. A particularly attractive case is to use high carrier frequencies where disturbing environment noise is lower and sensitivity of hearing to detect annoying anchor sources is also lower.

1. INTRODUCTION

An important property needed in typical virtual and augmented reality applications where the subject moves is to position his/her head in order to produce percepts related to the coordinates of the real or virtual environment. So far the existing general techniques for positioning, such as GPS, are limited from this point of view; they show problems in precision and robustness, particularly inside buildings, and they may not detect the orientation (direction of facing) of the subject. Examples of acoustic locating and positioning techniques are discussed for example in [1] and [2].

The present study builds on a future vision of *mobile* and *wearable augmented reality audio* (MARA, WARA) [2], based on subjects wearing binaural headsets with microphones close to ear canal entrances. In such cases there are binaural signals available which can be utilized to position the subject in relation to anchor sound sources [3]. Figure 1 characterizes a case where two (or four) loudspeakers (S1-S4) are the anchor sources, and the subject (R) wearing earplug microphones can move in the room.

Now the problem of acoustic estimation of position and orientation of the subject is related to robust estimation of the sound propagation times (or their differences) and levels at the ear microphones. In section 2 we discuss different conditions of obtaining enough information for reliable positioning of a subject.

The study reported in this paper is continuation to a previous work [3] where a relatively general framework was defined to estimate a subject's position using binaural microphone signals in different acoustic conditions. In this paper we focus on two specific ideas. First we utilize high audio frequencies radiated from anchor sources which brings some potential advantages compared to low and mid frequencies. Secondly we generate the anchor source signals in a way that makes them correlate after demodulation but not correlate as acoustic signals propagating in the air. System level details are discussed and studied, and experimental results from such a positioning system are presented.

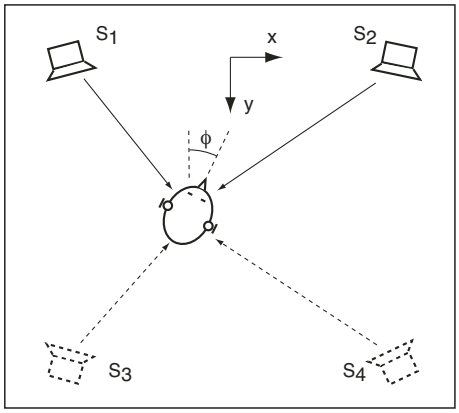

Fig. 1. An example of positioning system setup in a room.

2. SYSTEM CONSIDERATIONS

The basic goal in the case of a subject wearing binaural microphones in a room setup such as in Fig. 1 is to get a good estimate of the position and orientation, with as low disturbing effect as possible to the subject from the anchor source sounds.

Accuracy of position required depends on application but is often within 0.1-1 meters. Desired accuracy of estimated facing direction may vary between 5-20°. In special cases, such as scientific experiments, the required accuracy can be much higher.

Anchor sources of interest in this study are loudspeakers, although in a general case they may be any sound sources with some known or detectable characteristics [3]. The anchor sounds may be noise, signals with modulated data, music, etc. Because the detection techniques are typically based on cross-correlation of received signals, the periodicity of the anchor signals should be long enough so that the delays in a room can be estimated uniquely.

Binaural microphones are assumed to be close to the ears (preferably ear canal entrances) of the subject so that they capture a binaural signal which can be utilized in various ways in MARA systems [2]. From the positioning point of view the information obtainable consists of the distances to the anchor sources but also the interaural time and level differences (ITDs and ILDs) for orientation analysis. The binaural microphones must capture the frequency band of interest with a good enough S/N ratio.

Effects of **room acoustics** and **acoustic noise** are important to the robustness of positioning. Any deviation from a free space, such as reflection from surfaces of the room and objects inside the room, reverberation, diffraction and scattering, occlusion and shadowing of direct path by objects in between (including subject's head), brings complications to the estimation of features for positioning. Background noise means here everything else except the signals used for positioning, including subject's own speech that propagates effectively to the binaural microphones.

Perceptual factors from the viewpoint of a subject are the *noticeability* and *annoyance* of the anchor sound sources. In an ideal case the subject never perceives any anchor sound. This means that the level is below the masked threshold of hearing due to other sounds, or the anchor sound may be part of some desired sound such as reproduced music. Anchor sound level needs adjustment according to the level of background noise in the room.

Techniques for the **detection of anchor sounds** and analysis of position are dependent on the signalling conditions in the system. Signal processing may take place in a wearable unit (such as a pda device or a cellular phone) or it may be carried out elsewhere. The following two cases of anchor-related information are of interest here:

1. Anchor source signals are known in addition to the received binaural signals so that it is possible to estimate absolute distances to the sources. The minimum number of sources for unique positioning in the horizontal plane is three, not lying on the same line. With two sources, such as S_1 and S_2 in Fig. 1, unique positioning is possible within each half-plane divided by the line through the sources.
2. Anchor signals are not known but signals received can be cross-correlated to obtain the differences of distances between pairs of sources. The minimum number of sources for unique positioning in the horizontal plane is three.

Using a higher number of anchor sources than the minimum required helps to improve robustness because various techniques to find the optimally accurate/reliable solution are possible. Notice also that the anchor signals may carry other useful information than the one for positioning. They may be modulated by any data that fits to the transmission bandwidth used.

3. MODULATED HIGH-FREQUENCY SOURCES

In an earlier work on binaural positioning [3] we studied the estimation of anchor source distances by cross-correlation techniques using both wide-band and narrow-band (split-band) configurations. In the present study we introduce two details to the methodology: (1) high audio frequencies for anchor sources and (2) anchor sources produced from the same signal by modulation.

3.1. Usage of high-frequency anchor signals

High audio frequencies, such as 8–20 kHz, show some advantages as anchor sources compared to lower audio frequencies. Their main benefits are:

- *Background noise* level in rooms typically decreases toward high frequencies so that a good S/N ratio is possible with lower anchor signal levels, see Fig. 2(a) for typical office room noise.

- *Subject's own speech* is one of the strongest interfering signals to conflict with the anchor sounds. When high-frequency anchors are used, voiced sounds don't interfere as easily due to their low-pass behavior, see Fig. 2(b) for vowel /a:/ at a speaker's ear.

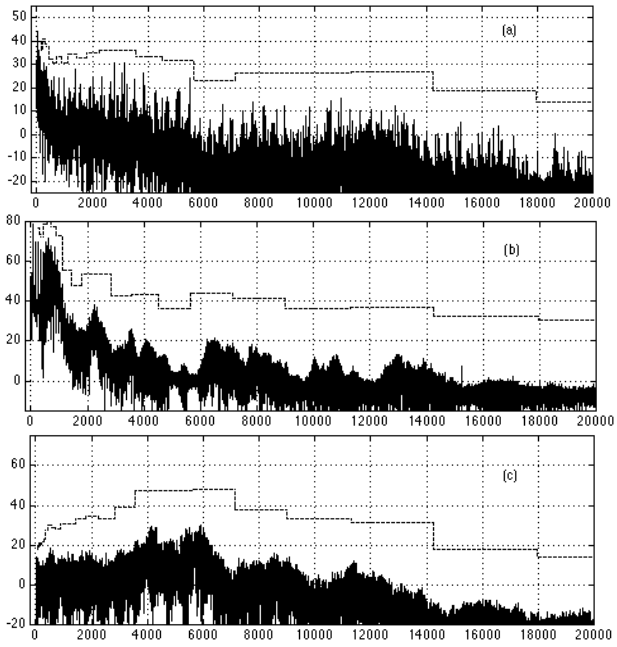

Fig. 2. Spectra of noises interfering with anchor sounds (linear frequency scale and no frequency weighting used for easy physical interpretation). (a) Background noise spectrum in a typical office room, consisting of ventilation noise (lowest frequencies) and computer fan noise. (b) Vowel /a:/ spectrum pronounced normally by a male subject, measured 1cm from ear canal entrance. (c) Fricative /s/ in vovel /a:/ context, same conditions as in case (b). Vertical axis: calibrated sound pressure level; horizontal axis: frequency in Hz. Dotted lines for 1/3-octave spectra. Signals were recorded through B&K 2238 sound level meter.

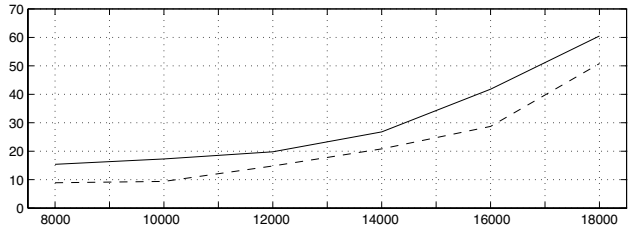

Fig. 3. JND threshold of anchor sound (Δf = 2 kHz) in dB as a function of modulation frequency (a) for bandpass noise and (b) bandpass-filtered impulse train (repetition period 100 ms). Average of two young subjects measured in an anechoic chamber.

Some fricatives may have stronger components at high frequencies, but due to diretivity of mouth radiation at highest frequencies, also they are attenuated at the ear canal entrance, see Fig. 2(c).

- *Auditory sensitivity* decreases at those frequencies so that anchor sounds of a given level are not as easily noticeable when the frequency is higher, as characterized in Fig. 3. Frequency masking also exhibits more spreading toward high frequencies, which further decreases the detectability of high-frequency anchor sounds.

- *Surface absorption* in a room typically increases at high frequencies, which means less reflections from walls and objects.

There are also drawbacks from using high audio frequencies, such as increased shadowing due to occluding objects, including

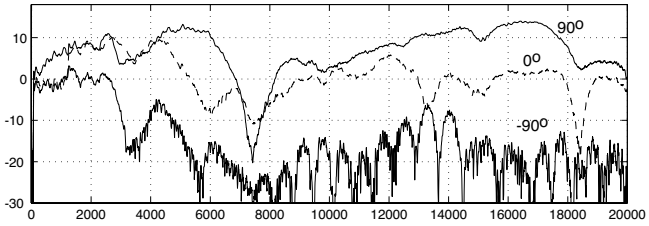

Fig. 4. Typical HRTF magnitude responses at ear canal entrance for sound arriving at 0° and $\pm 90^\circ$ angles in the horizontal plane.

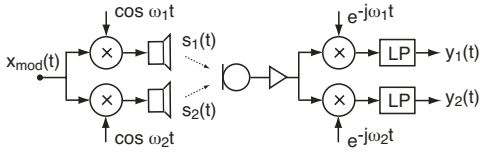

Fig. 5. Modulation/demodulation scheme for transmitting non-correlating acoustic anchor sounds but having common modulation. Two channels, received through one microphone, are shown.

subject's own head. This lowers the S/N ratio at binaural microphones in such cases. Figure 4 shows a typical case of HRTF magnitude responses for sound arriving to contralateral and ipsilateral ear ($\pm 90^\circ$) as well as to frontal sound incidence. It can be found that the attenuation due to shadowing for -90° is 10–25 dB at frequencies above 5 kHz.

Ultrasound ($f > 20$ kHz) is an attractive choice for anchors because then their levels can be much stronger without becoming noticeable, but the problem is that standard loudspeakers and microphones have rapidly decreasing sensitivity above 20 kHz.

One way to make the anchor sounds unnoticeable or less annoying is to use them as a part of informative or entertaining sounds. When music is reproduced to the room, anchor sounds can be embedded in it (possibly combined with watermarking). The anchor level must vary dynamically according to the masking level.

3.2. Coherently modulated anchor signals

Estimation of distances based on anchor sources requires that there is correlation between (a) the anchor and receiver sounds or (b) pairs of received sounds. If the anchor signals are known, condition (a) is easily met. Case (b) is, however, more complicated. The following strategy of using non-overlapping frequency bands with coherent modulation is found advantageous in both cases.

Figure 5 depicts the modulation/demodulation scheme applied here for anchor sounds. The original signal $x_{\mathrm{mod}}(t)$ can in principle be any low-pass signal, such as band-limited noise or a periodic signal with long enough period. The modulated anchor signal sent to loudspeaker i is obtained simply by

$$x_{i,\mathrm{anchor}}(t) = x_{\mathrm{mod}}(t)\ \cos(2\pi f_i t) \tag{1}$$

Demodulation is done by multiplication with complex-valued carrier followed by low-pass filtering. Complex-valued demodulation (by $e^{-j\omega_i t}$) and further cross-correlation processing are needed since the sidebands of the received acoustic signal are not mirror images. Frequency division into non-overlapping channels makes the anchor sounds non-correlating as acoustic signals but correlating after demodulation. Using a low-pass signal for modulation also helps reducing computation load of cross-correlation between signals by decimation of the demodulated signals.

Another, mathematically equivalent method of processing is to do most of the computation in the frequency domain through FFT and inverse FFT. In that case modulation is simply a shift in frequency, and cross-correlation is multiplication of one spectrum by the complex conjugate of another spectrum.

3.3. Estimation of position and orientation

After demodulation (and decimation) of signals, cross-correlation $R_{i,j}(\tau)$ between them to obtain an estimate of delay can be efficiently computed through the frequency domain:

$$r_{i,j}(\tau) = \int_{\tau-T}^{\tau} E\{Y_i(f)\, Y_j(f)^*\} e^{j2\pi f \tau} df \tag{2}$$

where $E\{Y_i(f)\ Y_j(f)^*\}$ is the expectation value of the cross-spectrum between signals $y_i(t)$ and $y_j(t)$. Now a simple estimate for delay between $y_i(t)$ and $y_j(t)$ is

$$d_{i,j} = \max_{\tau} \arg(|r_{i,j}(\tau)|) \tag{3}$$

In discrete-time computation (particularly for down-sampled signals) the exact position of the peak has to be interpolated, for example by parabolic fitting to three topmost samples. Another way to increase the accuracy of correlation peak detection is to use zero padding in the frequency domain processing of cross-correlation so that no peak position interpolation in the time domain is needed.

In ideal acoustic conditions the bandwidth of anchor signals does not have direct effect on positioning accuracy. However, with decreasing bandwidth the cross-correlation peak gets broader so that spatial resolution is reduced because of overlapping of correlation peaks of nearby objects. For example with modulation bandwidth of 1 kHz (2 kHz transmission bandwidth) the main lobe of cross-correlation function is 1 ms wide, corresponding to a distance of about 33 cm in wave propagation. A transmission band of at least 1-2 kHz is desirable to avoid the effect of falling into a narrow dip in acoustic transfer characteristics, see Fig. 4.

In practice there is need for several heuristics to improve robustness of positioning. Among features that can be analyzed based on cross-correlation are for example the peak-to-rms ratio $r_{i,j}(d_{i,j})/\mathrm{rms}(r_{i,j}(\tau))$, the ratio of the desired peak to other ones, etc. These can be utilized for instance to avoid abrupt jumps in distance estimate due to temporary noise bursts.

For estimating subject's position in the case 1 of Section 2, i.e., when anchor signals and their coordinates are known, the distances to left and right ear from an anchor can simply be averaged. If one of related cross-correlations is weak, only the stronger ear distance may be utilized. In case 2, i.e., when anchor signals are not available, only distance differences to anchor pairs can be obtained, separately for each ear.

Solving for the position of a subject's head when distances (or their differences) and anchor positions are known is a standard geometrical problem and is not discussed here in more detail. It is obvious that using more anchor sources makes the problem more overdetermined, which helps in improving robustness.

Estimation of subject's orientation angle ϕ (see Fig. 1) is based on interaural time difference (ITD) and possibly weakly on level difference (ILD), although the latter one is too irregular at highest audio frequencies to be a robust cue. ITD can be obtained from distance differences to the ears or separately by interaural signal cross-correlation through Eqs. (2)-(3). The facing angle φ_i in relation to anchor i can be resolved by assuming an ITD model, such as delay $d_{i,\mathrm{ITD}} = R_{\mathrm{head}}\{\varphi_i + \sin(\varphi_i)\}$, where R_{head} is the radius of subject's head. Front-back confusion can be resolved when two or more anchors are available.

4. EXPERIMENTS

In favorable acoustic conditions the accuracy of positioning and orientation can easily be good enough for practical applications, as studied and demonstrated earlier [3]. In the present work we concentrated on making the anchor sounds unnoticeable and studying the reliability of the approach in reverberant and noisy conditions.

From the experiment reported in Fig. 3 we can see that the JND threshold of anchor sound increases rapidly toward 20 kHz. Thus it is desirable to modulate the anchor sounds to as high frequencies as the responses of loudspeakers and binaural microphones allow. Otherwise it is difficult to find a low enough anchor level not audible but with good noise robustness. In experiments on the effects of room reflections reported below we used anchor carrier frequencies of 16–18 kHz with sound pressure levels of 35-50 dB, which makes positioning insensitive to most ambient noise, including subject's own speech. A loudspeaker (Genelec 1029A) as an anchor source and binaural microphones (Sennheiser KE 4-211-2 electret capsules) close to the ear canals of a dummy head (Cortex MK1) were used in a room with reflecting surfaces.

The effect of a strong wall reflection close to an ear microphone as well as a noticeable floor reflection is shown in Fig. 6(a). The direct path distance is slightly less than 3 meters and the strong reflection from a wall travels a 0.5 m longer path. In this case the peaks in the cross-correlation envelope are easily separated, the direct path corresponding to the first (strongest) peak.

Subplots 6(b) and (c) depict right and left ear cross-correlations with source when the right ear is facing to the source and the left ear is strongly shadowed by the head. In this case the right ear response is very clean, while direct sound to the left ear is almost 30 dB lower in level, and wall reflections coming later are much stronger. By relying on robust detection of direct sound by the right ear and searching for a correlation peak within 0.7 ms in left-ear cross-correlation, the weak peak can indicate the ITD delay.

ILD is so irregular at high frequencies that it is generally not useful except indicating on which side of the subject the source resides.

The worst case of delay estimation is when both ears are shadowed by an object between anchor source and binaural microphones, whereby wall reflections may be much stronger than the direct sound. This problem can be alleviated by using more ancor sources around the space where the subject moves. Detecting shadowed channels is then important by giving main emphasis to the most robust distance estimates. Sluggishness in reacting to rapid jumps in position or orientation helps and many other heuristic rules must be applied in a realistc full scale system.

5. SUMMARY AND DISCUSSION

In this study we have shown that acoustic positioning of a subject wearing binaural microphones in a limited space can be done by using anchor sound sources radiating high audio frequencies. By modulation and demodulation a low-pass anchor signal can be transmitted independently in several band-pass channels. It is demonstrated by experiments that cross-correlation between the received and the anchor signals yields distance from anchors to binaural microphones realiably in relatively severe acoustic conditions. Prototype tracker systems with 2 to 4 anchor source loudspeakers have been implemented to work in real time. A full-scale positioning system needs further heuristics to obtain the most robust estimate of subject's position in varying conditions, which is a problem outside the discussion scope of this study.

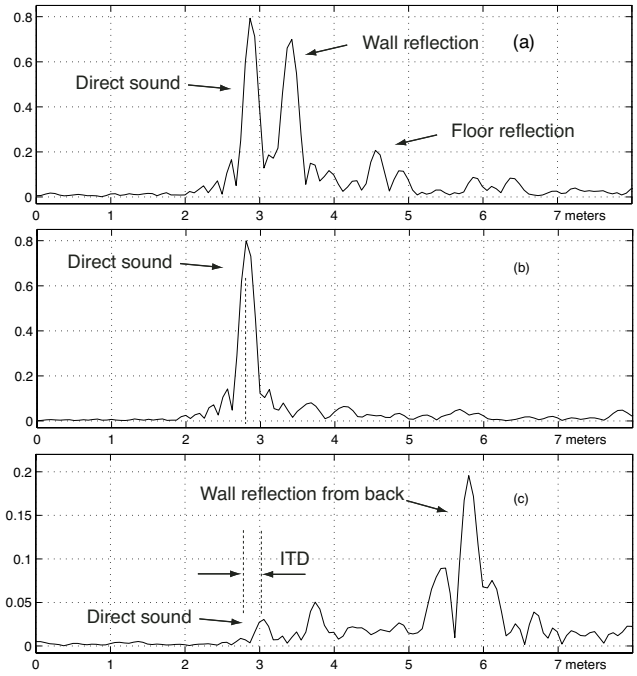

Fig. 6. Examples of cross-correlation envelopes in a real room. (a) Cross-correlation between anchor source and received signal near a reflective wall. (b) Right and (c) left ear cross-correlations when head-shadowing and back wall reflection are prominent in the left ear, while the right ear is directed towards the source. Horizontal axis: propagation delay mapped to distance in meters.

6. ACKNOWLEDGMENT

This work has been carried out in collaboration between Helsinki University of Technology and Nokia Research Center.

7. REFERENCES

[1] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust Localization in Reverberant Rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. Ward, Eds., Springer-Verlag, 2001, Ch. 7, pp. 131–154.

[2] A. Härmä, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki, H. Nironen, and S. Vesa, "Techniques and Applications of Wearable Augmented Reality Audio," in *Proc. 114th AES Convention*, preprint 5768, Amsterdam, March 22-25, 2003.

[3] M. Tikander, A. Härmä, and M. Karjalainen, "Binaural Positioning System for Wearable Augmented Reality Audio," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'03)*, New Paltz, N.Y., October 19-22, 2003.